Support Vector Machines + Kernel Methods

Machine Learning

Saket Anand

Source: [SS-Ch-12], [SS-Ch-15]

$$
\begin{aligned}
(x-y)^\top (x-y) &= x^\top x + y^\top y - x^\top y - y^\top x \\
&= K(x,x) + K(y,y) - 2K(x,y)
\end{aligned}
$$

SVM: Pimal and Dual Formulation

- Weak Duality

Primal Problem

$$\min_{\mathbf{w}} \max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \left(\frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right)$$

$$\geq \max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \min_{\mathbf{w}} \left(\frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right)$$

Dual Problem

- In our case, strong duality holds, i.e., we have equality
- So, the SVM objective is

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \min_{\mathbf{w}} \left(\frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right)$$

SVM: Dual Formulation

- SVM objective

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m:\boldsymbol{\alpha}\geq\mathbf{0}} \min_{\mathbf{w}} \left(\frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right)$$

- For fixed α, the problem wrt w is an unconstrained optimization problem, therefore the optimal w for a given α can be obtained from

$$\mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \; = \; 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

SVM: Dual Formulation

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m:\boldsymbol{\alpha}\geq\mathbf{0}} \left(\frac{1}{2}\left\| \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i \left\langle \sum_j \alpha_j y_j \mathbf{x}_j, \mathbf{x}_i \right\rangle \right) \right)$$

$$\boxed{\max_{\boldsymbol{\alpha}\in\mathbb{R}^m:\boldsymbol{\alpha}\geq\mathbf{0}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boxed{\langle \mathbf{x}_j, \mathbf{x}_i \rangle} \right)}$$

Only inner products

Kernel Methods: Motivation

- SVMs are binary classifiers that use hyperplanes to partition the data space
  - The data space may not be representative enough to have a *linear* boundary
- Example

Let the domain be the real line; consider the domain points $\{-10, -9, -8, \ldots, 0, 1, \ldots, 9, 10\}$ where the labels are $+1$ for all x such that $|x| > 2$ and -1 otherwise.

define a mapping $\psi : \mathbb{R} \to \mathbb{R}^2$ as follows:

$$\psi(x) = (x, x^2)$$

feature space to denote the range of ψ

Kernel Methods: Motivation

- Higher dimensional spaces are typically more representative
- Original training data

$$S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$$

- Transformed training data

$$\hat{S} = (\psi(\mathbf{x}_1), y_1), \ldots, (\psi(\mathbf{x}_m), y_m)$$

- How many dimensions should $\psi(\mathbf{x})$ have? What should be the functional form of $\psi(\mathbf{x})$?

Kernel Methods

- High dimensionality of $\psi(\mathbf{x})$ implies high computational cost
  - also needs more training data
- "Kernels" are used to define inner products in the feature space.

$$K(\mathbf{x}, \mathbf{x}') = \langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle$$

  - Can be interpreted as a similarity measure
- A different kind of representation

Standard vector representation with canonical bases

$$\boxed{\begin{aligned} \mathbf{x} &= x_1 \widehat{\mathbf{e}}_1 + x_2 \widehat{\mathbf{e}}_2 + \cdots + x_d \widehat{\mathbf{e}}_d \\ &= \sum_{i=1}^{d} \left(\mathbf{x}^\top \widehat{\mathbf{e}}_i\right) \widehat{\mathbf{e}}_i \end{aligned}}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \widehat{\mathbf{e}}_1 & \mathbf{x}_2^\top \widehat{\mathbf{e}}_1 & \ldots & \mathbf{x}_n^\top \widehat{\mathbf{e}}_1 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1^\top \widehat{\mathbf{e}}_n & \mathbf{x}_n^\top \widehat{\mathbf{e}}_n & \ldots & \mathbf{x}_n^\top \widehat{\mathbf{e}}_n \end{bmatrix}$$

Kernel Methods: The Kernel Trick

$$\mathbf{z}^\top \mathbf{K}\mathbf{z} = \mathbf{z}^\top \mathbf{X}^\top \mathbf{X}\mathbf{z}$$
$$= (\mathbf{X}\mathbf{z})^\top (\mathbf{X}\mathbf{z})$$
$$= \boxed{\widehat{\mathbf{z}}^\top \widehat{\mathbf{z}} > 0}$$

$$\mathbf{K} = \mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \mathbf{x}_1 & \mathbf{x}_1^\top \mathbf{x}_2 & \dots & \mathbf{x}_1^\top \mathbf{x}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n^\top \mathbf{x}_1 & \mathbf{x}_n^\top \mathbf{x}_2 & \dots & \mathbf{x}_n^\top \mathbf{x}_n \end{bmatrix}$$

Linear Kernel

where, $\widehat{\mathbf{z}} = \mathbf{X}\mathbf{z}$

quadratic term; always Positive for non-zero $\widehat{\mathbf{z}}$

Let $\mathbf{z} \in \mathbb{R}^n$ and $\mathbf{K} \in \mathbb{R}^{n\times n}$, then $\mathbf{K}$ is positive definite if and only if $\mathbf{z}^\top \mathbf{K}\mathbf{z} > 0, \forall \mathbf{z} \neq 0$.

- The kernel matrix is symmetric positive definite
- Any kernel function that induces such a matrix is a positive definite kernel function
- Also called as ‘Mercer Kernel’ (owing to Mercer’s theorem)

Kernel Methods: Example Kernel Function

$$K(x,z) = (x^T z)^2$$

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{n} x_i z_i\right)\left(\sum_{j=1}^{n} x_i z_i\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{n} (x_i x_j)(z_i z_j) \end{aligned}$$

$$\psi(\mathbf{x}) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

Induced feature transform

Example: Polynomial Kernel

$$
\begin{aligned}
K(x,z) &= (x^T z + c)^2 \\
&= \sum_{i,j=1}^{n} (x_i x_j)(z_i z_j) + \sum_{i=1}^{n} (\sqrt{2c}x_i)(\sqrt{2c}z_i) + c^2
\end{aligned}
$$

$$
\psi(\mathbf{x}) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \\ \sqrt{2c}x_1 \\ \sqrt{2c}x_2 \\ \sqrt{2c}x_3 \\ c \end{bmatrix}
$$

Induced feature transform

Gaussian Kernel

$$K(x, z) = \exp\left(-\frac{||x - z||^2}{2\sigma^2}\right)$$

- Also called as Radial Basis Function (RBF)
  - Infinite dimension feature space
  - Can vary the sigma parameter to change dimensionality of subspace induced by the kernel matrix

Kernel Methods

- Why is this representation important?
  - Gives us an *implicit* mapping to a feature space, i.e., without explicitly defining $\psi(\mathbf{x})$
  - Can choose kernels
    - with the *induced* feature spaces having arbitrarily high dimensionality
    - yet learn a classifier (e.g., SVM) in a 'subspace'
- The entire data is represented using a Kernel Matrix $\mathbf{K}$

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ K_{n1} & K_{n2} & \dots & K_{nn} \end{bmatrix} = \begin{bmatrix} \psi(\mathbf{x}_1)^\top \psi(\mathbf{x}_1) & \psi(\mathbf{x}_1)^\top \psi(\mathbf{x}_2) & \dots & \psi(\mathbf{x}_1)^\top \psi(\mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \psi(\mathbf{x}_n)^\top \psi(\mathbf{x}_1) & \psi(\mathbf{x}_n)^\top \psi(\mathbf{x}_2) & \dots & \psi(\mathbf{x}_n)^\top \psi(\mathbf{x}_n) \end{bmatrix}$$

SVM: Dual Formulation with Kernels

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ K_{n1} & K_{n2} & \dots & K_{nn} \end{bmatrix} = \begin{bmatrix} \psi(\mathbf{x}_1)^\top \psi(\mathbf{x}_1) & \psi(\mathbf{x}_1)^\top \psi(\mathbf{x}_2) & \dots & \psi(\mathbf{x}_1)^\top \psi(\mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \psi(\mathbf{x}_n)^\top \psi(\mathbf{x}_1) & \psi(\mathbf{x}_n)^\top \psi(\mathbf{x}_2) & \dots & \psi(\mathbf{x}_n)^\top \psi(\mathbf{x}_n) \end{bmatrix}$$

$$\boxed{\max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boxed{\langle \mathbf{x}_j, \mathbf{x}_i \rangle} \right)}$$

Only inner products

replace

$$\boxed{\langle \psi(\mathbf{x}_j), \psi(\mathbf{x}_i) \rangle = K(\mathbf{x}_j, \mathbf{x}_i)}$$